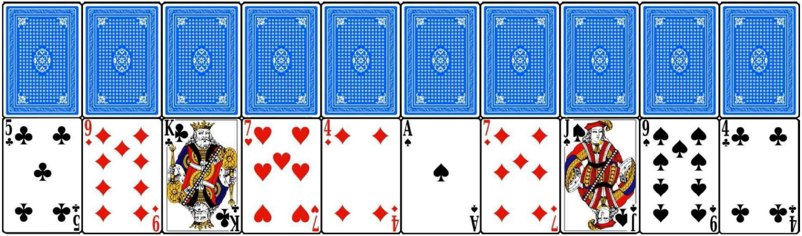
Прикладная статистика
4. Множественная проверка гипотез, последовательный анализ.

Родионов Игорь Владимирович
rodionov@bigdatateam.org

Поиск экстрасенсов

В рамках исследования возможностей экстрасенсорного восприятия осуществлялся поиск экстрасенсов. Испытуемому предлагалось угадать цвет 10 карт.



H_0 : испытуемый выбирает ответ наугад;
H_1 : испытуемый может предсказывать цвета карт.
Статистика t – число карт, цвета которых угаданы.

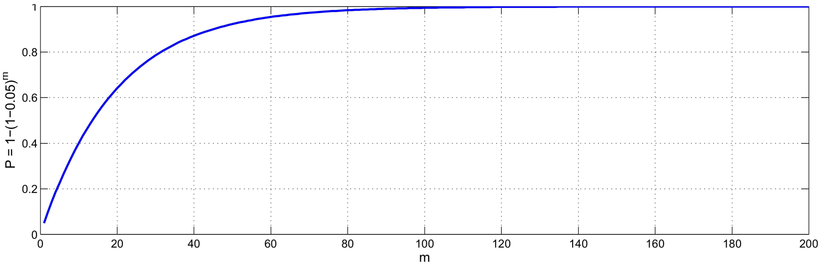
$$P(t \geq 9|H_0) = P(t = 9|H_0) + P(t = 10|H_0) = 10 \cdot (0.5)^{10} + (0.5)^{10} \approx 0.01,$$

т.е. при $t \geq 9$ можно отвергать H_0 на уровне значимости 0.02.

Поиск экстрасенсов

Процедуру отбора прошли 1000 человек.
Девять из них угадали цвета 9 из 10 карт, двое – цвета всех 10 карт.
Ни один в последующих экспериментах не подтвердил своих способностей.

Вероятность того, что из 1000 человек хотя бы один случайно угадает цвета 9 или 10 из 10 карт равна $(1 - 11 \cdot (0.5)^{10})^{1000} \approx 0.99998$.



Постановка задачи

Пусть имеются данные $X = \{X_i^{(j)}\}$, $1 \le i \le n_j$, $1 \le j \le m$.

По ним проверяем гипотезы $H_j : P_j \in \mathcal{P}_j$ против альтернатив $H'_j : P_j \notin \mathcal{P}_j$ с помощью статистик $T_j = T_j(X_1^{(j)}, \ldots, X_{n_j}^{(j)})$. Пусть $p_j = p_j(T_j)$ – p-значения критериев.

Обозначим $M = \{1, \ldots, m\}$, M_0 – индексы верных гипотез, $|M_0| = m_0$, R – число отвергнутых гипотез, V – число ошибок первого рода.

	$\sharp$ верных H_j	$\sharp$ ложных H_j	Всего
$\sharp$ принятых H_j	U	T	m-R
$\sharp$ отвергнутых H_j	V	S	R
Всего	m_0	$m-m_0$	m

Групповая вероятность ошибки I рода (family-wise error rate)

$$FWER = \mathrm{P}(V > 0).$$

Контроль над FWER на уровне α означает, что

$$FWER = P(V > 0) \leq \alpha$$

для всех распределений из верных гипотез H_j, $j \in M_0$.

Пусть $\alpha_1, \ldots, \alpha_m$ – уровни значимости критериев проверки гипотез $H_1, \ldots, H_m$ соответственно. Хотим их выбрать таким образом, чтобы $FWER \leq \alpha$.

Метод Бонферрони: $\alpha_1 = \ldots = \alpha_m = \frac{\alpha}{m}$.
Действительно,

$$FWER = P(V > 0) = P(\exists j \in M_0 : \ p_j \leq \alpha/m) \leq$$

$$\sum_{j \in M_0} P(p_j \leq \alpha/m) \leq m_0 \cdot \frac{\alpha}{m} \leq \alpha.$$

Главный недостаток метода – резкое уменьшение мощности статистической процедуры при $m \to \infty$.

Пример: критерий Стьюдента для независимых выборок, $(X_1^1, \dots, X_n^1)$ – выборка из $N(\mu_1, 1)$, $(X_1^2, \dots, X_n^2)$ – выборка из $N(\mu_2, 1)$, $\mu_2 - \mu_1 = 1$, повторяем эксперимент m раз, $H_0 : \mu_1 = \mu_2$, $H_1 : \mu_1 \neq \mu_2$.

m	n	Мощность
1	23	0.9
10	23	0.67
100	23	0.37
1000	23	0.16

Метод Шидака: $\alpha_1 \dots \alpha_m = 1 - (1-\alpha)^{1/m}$.

Метод дает $FWER \leq \alpha$ при условии, что статистики T_i независимы или выполнено свойство "положительной зависимости":

$$P(T_1 \leq t_1, \dots, T_m \leq t_m) \geq \prod_{i=1}^{m} P(T_i \leq t_i) \ \forall \vec{t} \in \mathbb{R}^m.$$

Положительную зависимость, в частности, можно установить с помощью FKG-неравенства: если $f(x)$ и $g(x)$ – возрастающие (убывающие) функции, то $Ef(X)g(X) \geq Ef(X)Eg(X)$.

Нисходящие процедуры

Составим вариационный ряд р-значений

$$p_{(1)} \leq \ldots \leq p_{(m)},$$

где $H_{(1)}, \ldots, H_{(m)}$ – соответствующие гипотезы. Процедура выглядит так:

1. Если $p_{(1)} \geq \alpha_1$, то принимаем все гипотезы $H_{(1)}, \ldots, H_{(m)}$ и останавливаемся, иначе отвергаем $H_{(1)}$ и продолжаем;
2. Если $p_{(2)} \geq \alpha_2$, то принимаем все гипотезы $H_{(2)}, \ldots, H_{(m)}$ и останавливаемся, иначе отвергаем $H_{(2)}$ и продолжаем;
3. $\ldots$

Нисходящие процедуры

Метод Холма: нисходящая процедура с уровнями значимости

$$\alpha_1 = \frac{\alpha}{m}, \dots, \alpha_i = \frac{\alpha}{m - i + 1}, \dots, \alpha_m = \alpha.$$

Свойства:

1. контролирует FWER на уровне значимости α;
2. равномерно мощнее метода Бонферрони;
3. если характер зависимости между статистиками $\{T_i\}$ неизвестен, то нельзя построить контролирующую FWER на уровне α процедуру мощнее, чем метод Холма.

Метод Шидака-Холма: нисходящая процедура с уровнями значимости

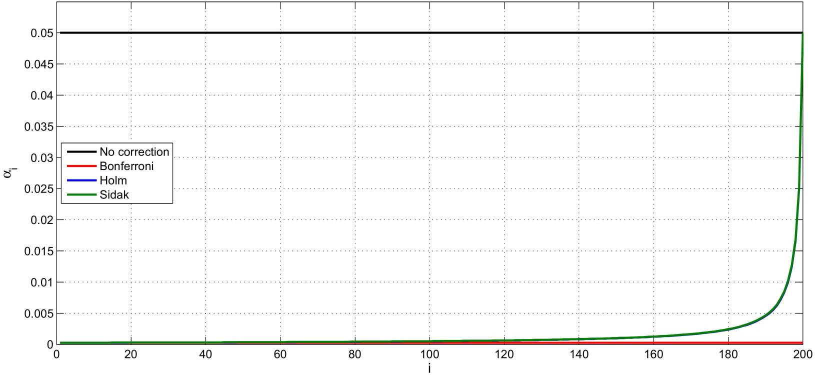
$$\alpha_1 = 1 - (1-\alpha)^{\frac{1}{m}}, \dots \alpha_i = 1 - (1-\alpha)^{\frac{1}{m-i+1}}, \dots \alpha_m = \alpha.$$

Свойства:

1. контролирует FWER на уровне значимости α, если статистики $\{T_i\}$ **независимы в совокупности**;
2. если статистики $\{T_i\}$ независимы в совокупности, то нельзя построить контролирующую FWER на уровне α процедуру мощнее, чем метод Шидака-Холма;
3. при больших m мало отличается от метода Холма.

Сравнение методов

На практике при больших m методы Холма и Шидака-Холма практически совпадают и являются более мощными, чем метод Бонферрони.



Ожидаемая доля ложных отклонений гипотез (false discovery rate)

$$FDR = \mathsf{E}\left(\frac{V}{\max(R, 1)}\right).$$

Контроль над FDR на уровне значимости α означает, что $FDR \leq \alpha$ для всех распределений из верных гипотез H_j, $j \in M_0$.

Хотя $FDR = \mathsf{E}\left(\frac{V}{\max(R,1)}\right) \leq \mathsf{E}I(V > 0) = P(V > 0) = FWER$, но в рамках процедур, контролирующих FDR на уровне α, случается больше ошибок первого рода.

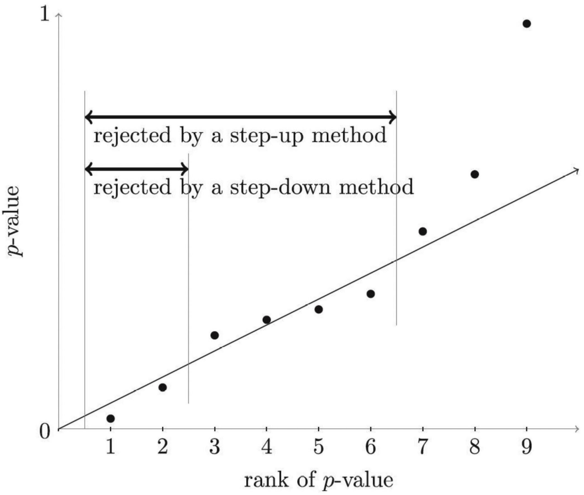
Восходящие процедуры

Пусть, как и ранее, $p_{(1)} \leq \ldots \leq p_{(m)}$ – вариационный ряд полученных p-значений, а $H_{(1)}, \ldots, H_{(m)}$ – соответствующие им гипотезы.

Процедура выглядит так:

1. Если $p_{(m)} < \alpha_m$, то отвергаем все гипотезы $H_{(1)}, \ldots, H_{(m)}$ и останавливаемся, иначе принимаем $H_{(m)}$ и продолжаем;
2. Если $p_{(m-1)} < \alpha_{m-1}$, то отвергаем все гипотезы $H_{(1)}, \ldots, H_{(m-1)}$ и останавливаемся, иначе принимаем $H_{(2)}$ и продолжаем;
3. ...

Восходящая процедура отвергает не меньше гипотез, чем нисходящая с теми же $\{p_i\}$ и $\{\alpha_i\}$.



Восходящие процедуры

Метод Бенджамини-Хохберга: восходящая процедура,

$$\text{для которой } \alpha_i = \alpha \cdot \frac{i}{m}, \;\; i = 1, \ldots, m.$$

Метод контролирует FDR на уровне α, если $\{T_i\}$ независимы или выполнено свойство PDRS:

$$P(X \in D | T_i = x) \text{ не убывает по } x \; \forall i \in M_0,$$

где D – возрастающее множество, т.е. если $\vec{y} \in D$ и $\vec{z} \geq \vec{y}$, то $\vec{z} \in D$.

В частности, свойство PDRS выполнено, если $X \sim N(a, \Sigma)$, где все элементы ковариационной матрицы Σ неотрицательны.

Восходящие процедуры

Метод Бенджамини-Иекутиели: восходящая процедура с уровнями значимости

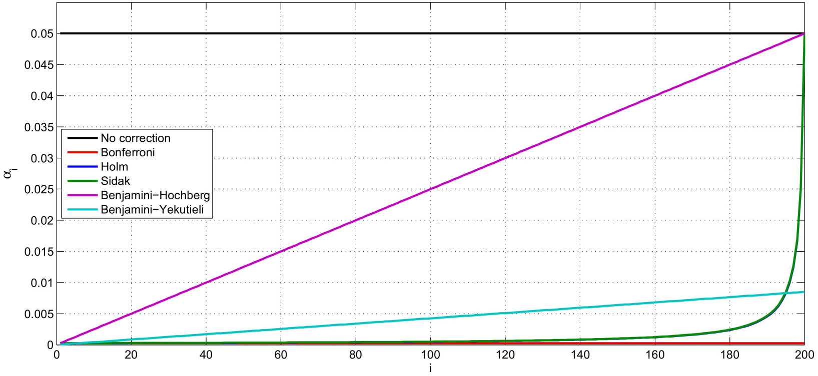
$$\alpha_i = \alpha \cdot \frac{i}{m} \left(\sum_{j=1}^{m} \frac{1}{j} \right)^{-1}, \quad i = 1, \dots, m.$$

Метод контролирует FDR на уровне $\frac{m_0}{m}\alpha \leq \alpha$ для любых T_i.

При отсутствии информации о зависимости между статистиками T_i метод не улучшаем.

Если доля неверных гипотез мала, то метод Бенджамини-Иекутиели отвергает меньше гипотез, чем метод Холма.

Сравнение методов



Модельный эксперимент: пусть имеется $m = 200$ выборок размера $n = 20$ из нормального распределения $N(a, 1)$, причем первые $m_0 = 150$ выборок сделаны из $N(0, 1)$, а последние 50 – из $N(1, 1)$.

Проверим гипотезы $H_i : a = 0$ против альтернатив $H_i' : a \neq 0$, $i = 1, \ldots, m$, с помощью t-критерия Стьюдента, полагая $\alpha = 0.1$:

$$\text{если } \left|\sqrt{n}\frac{\overline{X} - a}{s}\right| > t_{1-\alpha/2}, \text{ то отвергнуть } H_i,$$

где $t_{1-\alpha/2}$ – $(1 - \alpha/2)$-квантиль распределения Стьюдента $St(n-1)$.

Матрицы ошибок модельного эксперимента:

Без поправки

	True	*False*
Accepted	142	0
Rejected	8	50

Бонферрони

	True	*False*
Accepted	150	27
Rejected	0	23

Шидак-Холм

	True	*False*
Accepted	150	24
Rejected	0	26

Бенджамини-Хохберг

	True	*False*
Accepted	148	4
Rejected	2	46

Выводы

1. Если мы проверяем цепочку гипотез о каком-то одном наборе данных, то при отклонении одной из гипотез в рамках процедуры множественной проверки стоит остановиться и отклонить все остальные. Например, в модельном эксперименте стоит отвергнуть гипотезу о том, что данные выбраны из стандартного нормального распределения.
2. Если мы последовательно проверяем гипотезы о различных наборах данных, то процедура множественной проверки гипотез также необходима, поскольку если поправки не делать, то вероятность того, что произойдет ошибка первого рода, будет расти с количеством проверяемых гипотез.

Последовательный анализ

Пусть имеются однородные данные $(X_1, X_2, \ldots)$, поступающие с течением времени. Пусть $\forall i$ $X_i \sim P \in \{P_\theta, \theta \in \Theta\}$.

Хотим проверить простую гипотезу $H_0 : \theta = \theta_0$ против альтернативы $H_1 : \theta = \theta_1$ (если $\theta_1 > \theta_0$, то часто можно свести задачу к проверке гипотезы $H_0 : \theta \leq \theta_0$ против альтернативы $H_1 : \theta \geq \theta_1$).

Метод последовательного анализа позволяет существенно сократить количество наблюдений, необходимых для значимой процедуры проверки гипотез (до двух раз).

Последовательный анализ

Пусть $p_\theta(x)$ – плотность распределения P_θ (или вероятность $P_\theta(X_1 = x)$ в дискретном случае). Определим функцию правдоподобия

$$f_n(X, \theta) = p_\theta(X_1) \cdot \ldots \cdot p_\theta(X_n).$$

Выберем две константы A и B, $A > B$. На каждом шаге процедуры вычисляется отношение $R_m = \frac{f_m(X, \theta_1)}{f_m(X, \theta_0)}$ и

- если $R_m \geq A$, то отклоняем H_0;
- если $R_m \leq B$, то принимаем H_0;
- если $B < R_m < A$, то переходим к рассмотрению R_{m+1}.

Последовательный анализ

Как выбирать нижнюю и верхнюю границу в последовательном анализе? Предположим, мы хотим построить критерий уровня значимости α и мощности не менее β. Тогда можно выбрать

$$A = \frac{1-\beta}{\alpha}, \qquad B = \frac{\beta}{1-\alpha}.$$

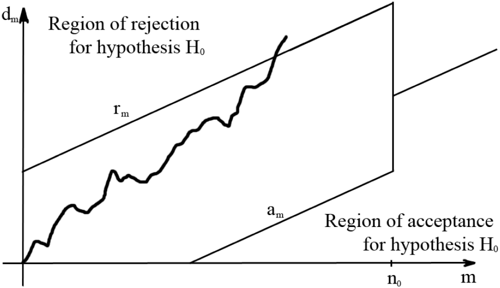
Кроме того, удобнее работать с $\ln R_m$, поскольку

$$\ln R_m = \sum_{i=1}^{m} \ln \frac{p_{\theta_1}(X_i)}{p_{\theta_0}(X_i)} = \sum_{i=1}^{m} Z_i,$$

т.е. мы получаем так называемое случайное блуждание, поведение которого просто изучать.

Усечение

В случае, если данных у нас ограниченное количество (n наблюдений), можно воспользоваться процедурой усечения: если до n-го момента мы не приняли решение, то если $R_n \geq \frac{A+B}{2}$, то отвергаем H_0, а если $R_n < \frac{A+B}{2}$, то принимаем H_0.



Данная процедура несколько снижает мощность критерия, но при больших n изменение не существенно.

Момент остановки

Пусть истинное значение параметра – θ. Момент остановки процедуры последовательного анализа - это случайная величина, найдем её среднее. Определим $h(\theta)$ как решение уравнения

$$\int_{\mathbb{R}} \left(\frac{p_{\theta_1}(x)}{p_{\theta_1}(x)} \right)^{h(\theta)} p_\theta(x) dx = 1.$$

Тогда среднее приблизительно равно

$$E_\theta n \approx \frac{L(\theta) \ln B + (1 - L(\theta)) \ln A}{E_\theta Z_1},$$

где

$$L(\theta) = \frac{A^{h(\theta)} - 1}{A^{h(\theta)} - B^{h(\theta)}}.$$

Задача: рекламная кампания планировалась так, чтобы обеспечить узнаваемость продукта среди целевой аудитории более 30%. После окончания кампании проводится опрос с целью оценки узнаваемости.

H_0: узнаваемость продукта не превышает 30%.
H_1: узнаваемость продукта превышает 30%.

Т.е. нам поступают случайные величины $(X_1, X_2, \ldots)$, распределенные по закону $Bern(\theta)$. В последовательном анализе необходим "зазор" между гипотезами, поэтому будем проверять гипотезу $H_0 : \theta < p_L = 0.3 - \delta$ против $H_1 : \theta > p_U = 0.3 + \delta$.

Пример

Рассмотрим статистику $d_m = \sum_{i=1}^m X_i$ и две границы

$$a_m = \frac{\ln B + m \ln \frac{1-p_L}{1-p_U}}{\ln \frac{p_U}{p_L} - \ln \frac{1-p_U}{1-p_L}}, \qquad r_m = \frac{\ln A + m \ln \frac{1-p_L}{1-p_U}}{\ln \frac{p_U}{p_L} - \ln \frac{1-p_U}{1-p_L}},$$

данные формулы получаются после раскрытия $\ln R_m$. Тогда процедура последовательного анализа будет выглядеть так: при каждом значении m

- если $d_m \geq r_m$, то отвергаем H_0, т.е. $\theta \geq p_U$;
- если $d_m \leq a_m$, то принимаем H_0, т.е. $\theta \leq p_L$;
- если $a_m < r_m < d_m$, то продолжаем процедуру и добавляем элемент выборки.

Спасибо за внимание!